

本周周报（2013.12.23-2013.12.29）

郭方舟

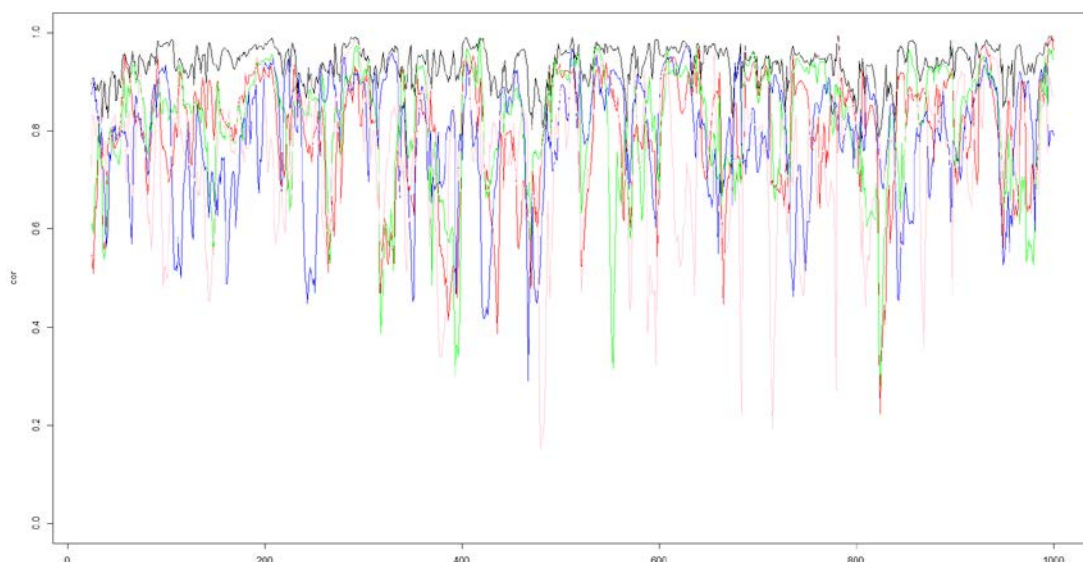
本周工作

1 空气污染数据可视化

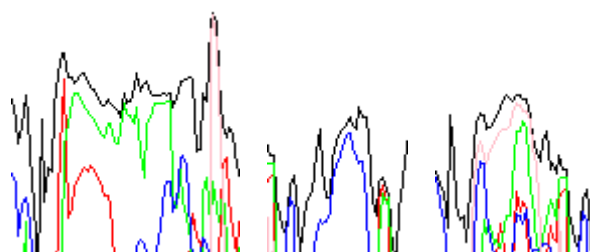
本周主要是在复习托福和完成作业，因此在这个任务上投入的时间不多。

CCA:

在计算 CCA 的时候，将属性分为两大类，污染属性和气象属性，其中，污染属性包含 CO，NO2，PM10，SO2，NOX 五个属性；气象属性包含 WS（风速）、WD（风向）、TEMP（温度）、HUMD（湿度）、PRESS（压力）五个属性。由于 WD 是一个矢量，因此在计算时没有包含。进行 CCA 分析后取第一典型相关向量对应的相关系数作为两个数据间的相关性。以 24 小时为窗长，窗移取 1 进行 CCA。总共计算了五组变量之间的相关性，五组变量分别为：污染属性与气象属性（黑线）、污染属性与风速（蓝线）、污染属性与温度（绿线）、污染属性与湿度（红线）和污染属性与压力（粉线）。得到的相关性如下图所示：



从图中观察到了一些有趣的现象，例如：



可以看到，在这三张截图中，某个彩色的线与黑色的线非常接近或者是走势一致，所以我们能不能根据 CCA 的计算结果看出气象数据中在某个时刻或时段内是哪一个或哪几个变量影响着污染数据，再通过可视化把这种信息表达出来。

分段线性回归度量相似度:

本周主要在看一些论文，找 cca 的例子以及时序序列的相似度度量方法。

有一个比较有趣的方法是基于分段线性逼近和 ddtw 的相似性度量方法。这种方法首先使用分段线性逼近提取时序序列的特征，再使用 ddtw 计算时序序列计算相似度。

与 sax 方法相比，使用分段线性逼近对数据进行处理的好处在于可以减少数据特征的丢失。不过计算两条序列的相似度这件事情我们还不知道需不需要，目前只是看看相关的论文，做做准备。

下周工作

1. 空气污染数据的可视化

找 CCA 相关的例子进行学习和实现。

找时序回归分析的论文和例子进行学习和实现

2. 考试复习

马上到考试周，有三门考试需要复习。